Same Analysis, Many Sources

Four patterns for RWE harmonization, and how they compose

One cohort study. Many sources. Four ways to implement it.

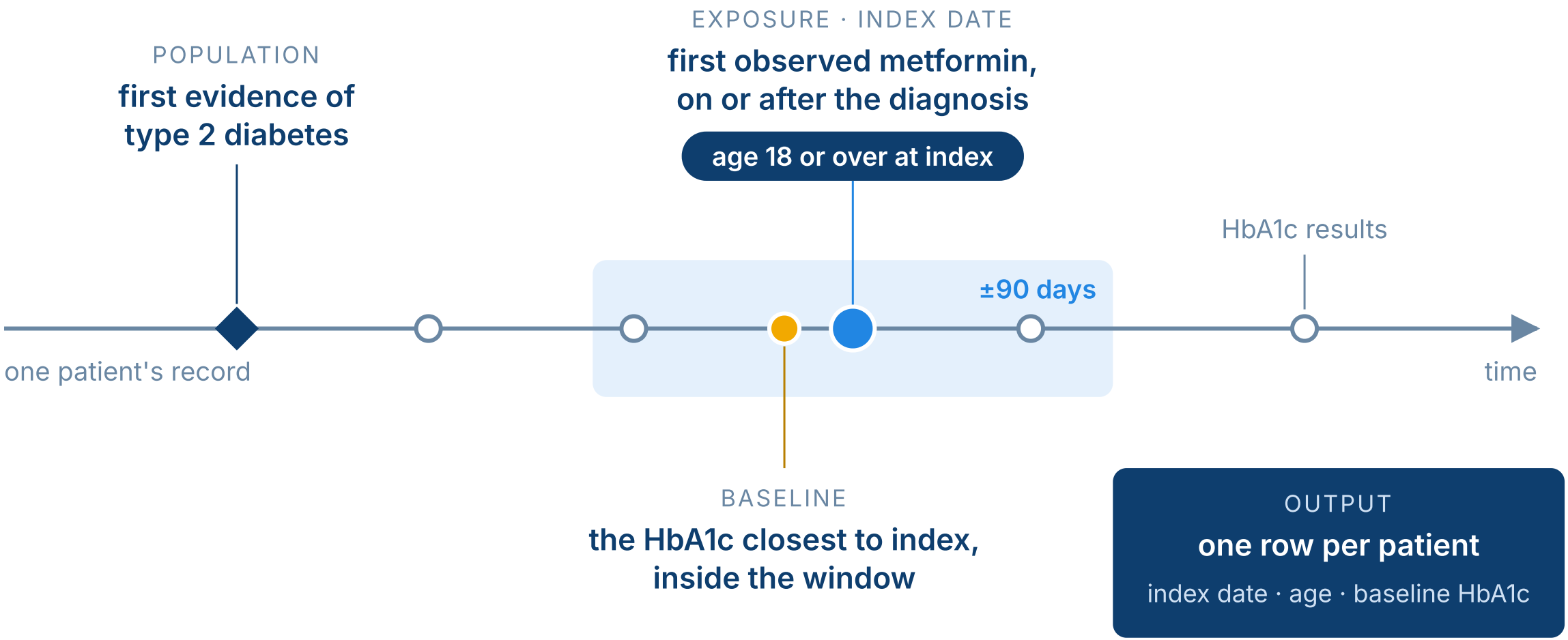
KIERAN MACE, PHD · PLINTH ANALYTICS · PHUSE REAL WORLD DATA AUTUMN EVENT · 1 OCTOBER 2026

Slides and demo code were developed with AI assistance (Claude); the study design, data, results and claims were reviewed by the author.

In the next 25 minutes

1. **One study**, written once as a protocol
2. **Three kinds of work** every implementation has to do
3. **Four places** to do that work, each run on the same three sources
4. **What differed**, why, and how to choose

An example study



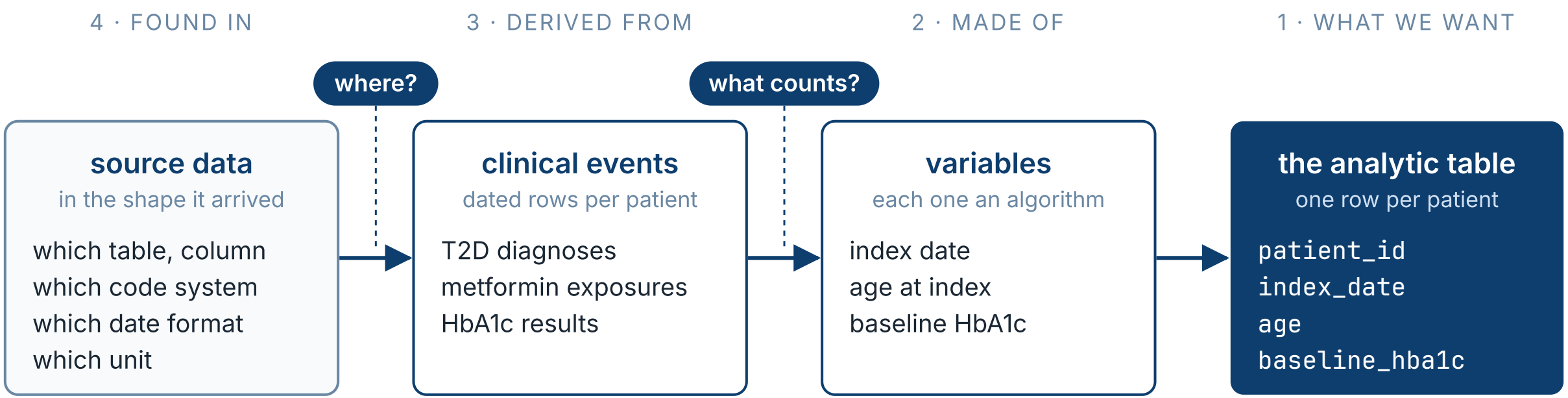
The study as code: the protocol

```
run_protocol <- function(dataset, window = c(-90, 90), mapping = NULL, ...) {
  dataset |>
    add_diagnoses(t2d, mapping = mapping, ...) |>                # find T2D
    add_exposures(metformin, mapping = mapping, ...) |>          # find metformin
    add_index_date(exposure = metformin, on_or_after = t2d) |>   # first metformin after T2D
    add_age(at = "index") |>
    restrict(age >= 18) |>
    add_baseline(hba1c, window = window, mapping = mapping) |>   # closest HbA1c
    pull_cohort()                                                # one row per patient
}
```

Three verbs **find evidence** in the source: a diagnosis, an exposure, a lab. The other four reason about dates and ages.

This block does not change for the rest of the talk. What changes is **what sits under the three evidence verbs**.

How an analytic table gets made



Where does the evidence live?
A data engineer answers it from the codebook.
Different for every source.

What counts as evidence?
A scientist decides, and defends it.
The same decision for every source.

An example phenotype: exposure to metformin

one row of the claims source, table **`prescription_drug_events`**

DESYNPUF_ID	SRVC_DT	PROD_SRVC_ID	DAYS_SUPLY_NUM
0A6F99CACEF0…	20091218	55289038494	30

1

Navigate the schema

Which table, which columns, and a date stored as text.

a codebook question

2

Interpret the values

A drug code in the NDC vocabulary: this one is metformin 500 mg.

a vocabulary question

3

Understand what a record means

A pharmacy fill. Is one fill exposure, or does it take a refill?

a scientific question

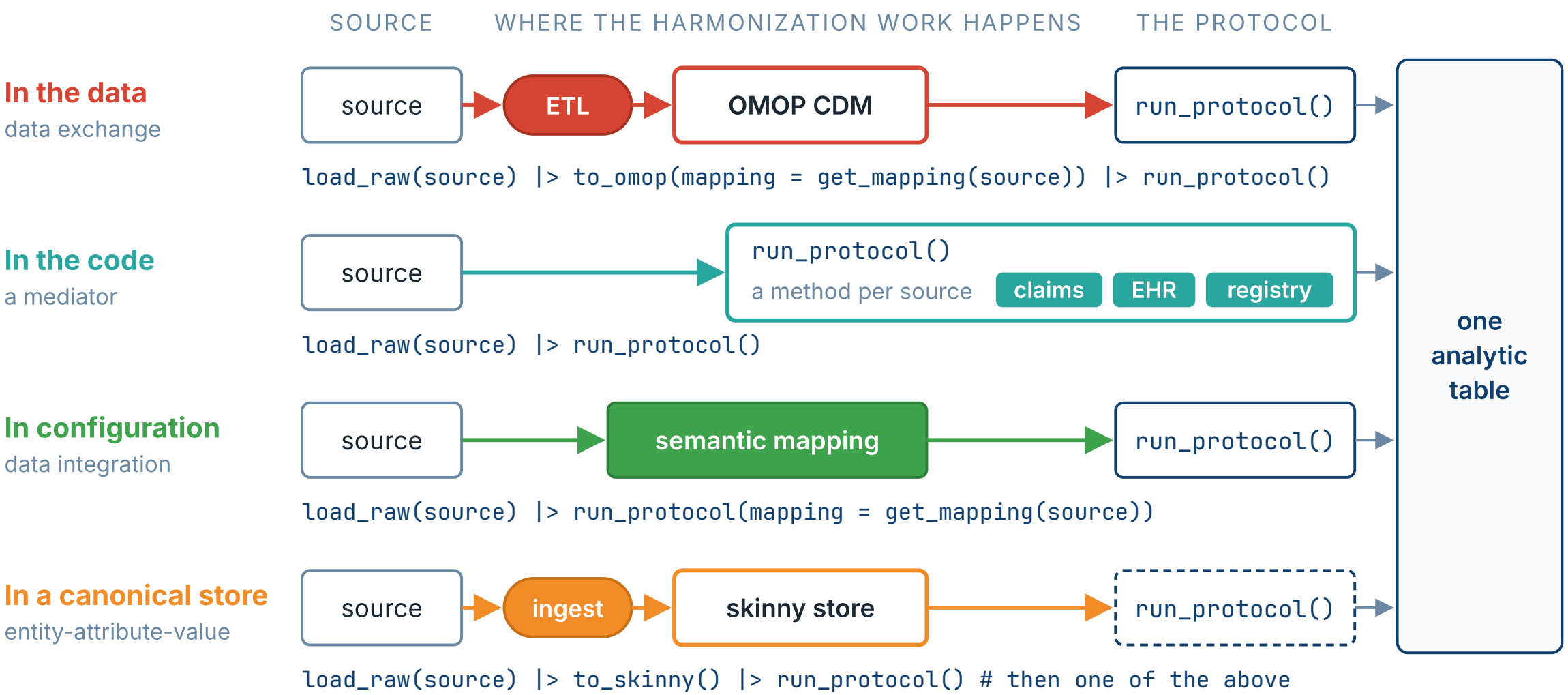
Every technique has to do all three. They differ in **which of the three they take on**, and **where the work sits**.

Database terminology

Data engineers have a formal name for each of the four places. The talk uses both.

In this talk	Database term	What it means
in the data	**data exchange**	convert the source once; every later query runs on the converted copy
in the code	**mediator, wrappers**	source-specific code behind one common interface
in configuration	**data integration**	leave the source as it is; translate each query when it runs
a canonical store	**entity-attribute-value**	one flat shape for everything; the meaning moves into the query

Three places to do the harmonization work, and a store that strips the schema first



plinth DATA ELEVATED

What each technique takes on

	in the data	in the code	in configuration	in a canonical store
1 Navigate the schema	**yes** the whole source, once	**yes** per source, per verb	**yes** the tables the protocol needs	**partly** one shape for every source
2 Interpret the values	**yes** every code, once	**yes** in the source's own vocabulary	**yes** the file names the vocabulary	**no** kept as they arrived
3 Understand what a record means	**recorded** but not decided	**yes** a rule in the method	**no** stays in the code	**no** left to every query

Level 3 is the science, and only the code takes it on.

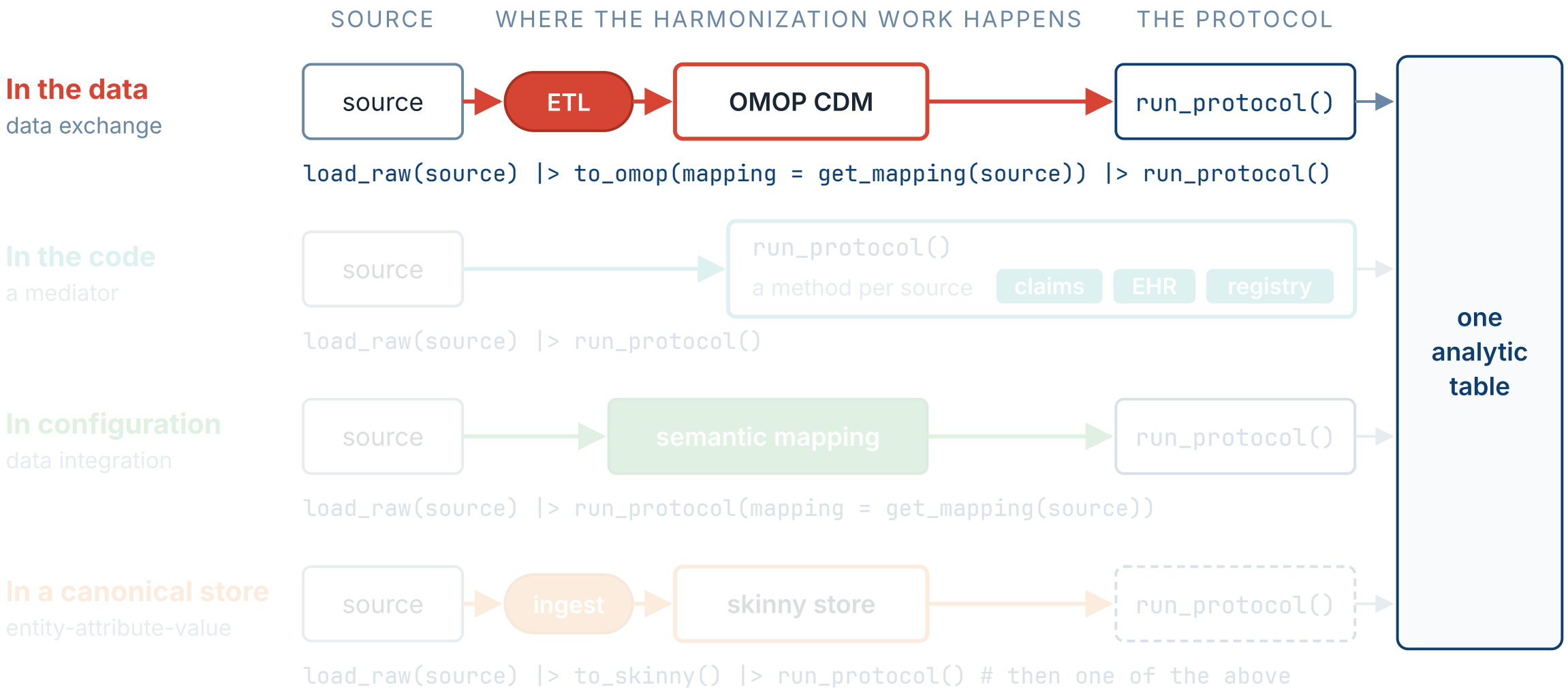
plinth DATA ELEVATED

The worked example: three sources, three phenotypes, nine operational definitions

Evidence	SynPUF claims	Synthea EHR	Registry extract
T2D diagnosis	ICD-9 250.x, in any of 10 columns	SNOMED 44054006	`diabetes_type = 'Type 2'`
Metformin exposure	an NDC on a Part D event	RxNorm 860975 on an order	`'metformin'` inside a string
What a drug row is	a **dispensing**	an **order**	a **line of therapy**
Baseline HbA1c	**not observable**	LOINC 4548-4	`lab_name = 'HbA1c'`

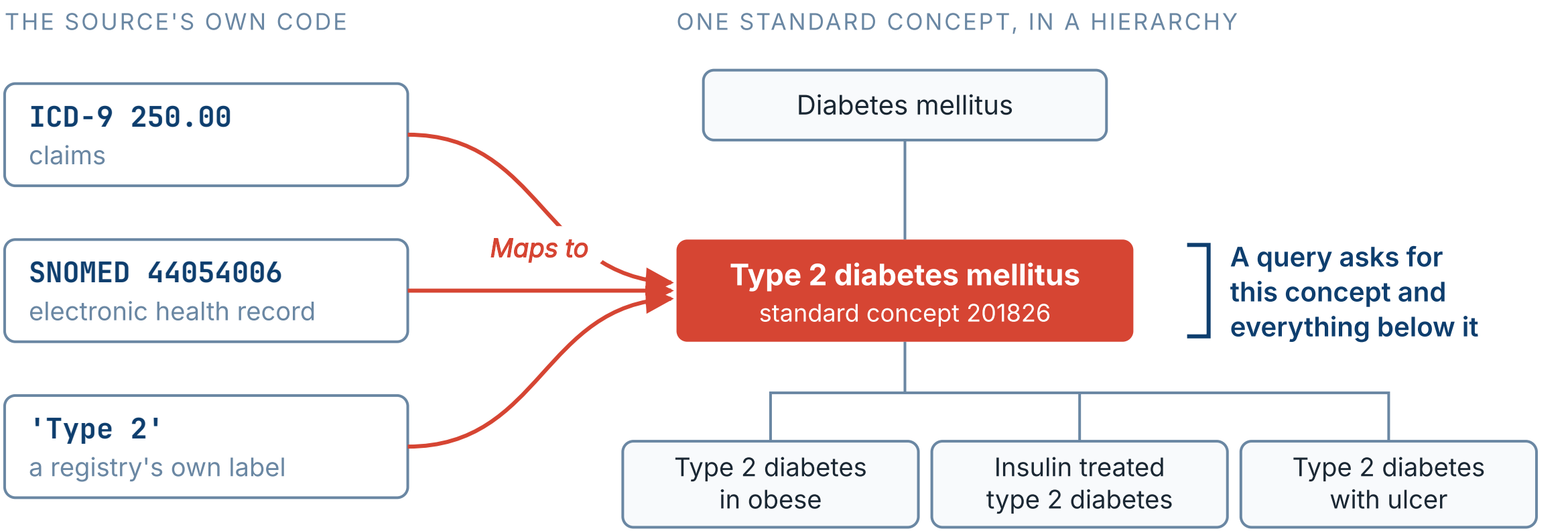
Three public, synthetic sources with **no shared schema, vocabulary or calendar**.

In the data: data exchange. Materialize the target once, upstream, for every query you might ask



plinth DATA ELEVATED

The OMOP common data model: one schema and one vocabulary for every source



One schema too: four tables hold every source's patients, diagnoses, drug exposures and lab results.

After the ETL the protocol has nothing to decide, and the claims cohort is 1,585 patients shorter

THE CLAIMS CODE		WHERE THE MAPPING FILED IT
250.00	no complication	**Type 2 diabetes mellitus** — what the standardized query asks for
250.40	with renal manifestations	Disorder of kidney due to diabetes mellitus
250.50	with eye manifestations	Disorder of eye due to diabetes mellitus
250.60	with nerve manifestations	Disorder of nervous system due to diabetes

4 of 51 codes land inside

47 land outside: kept in the CDM, never read by the query

Claims cohort **8,547** in the code → **6,962** over the CDM **1,585 patients fewer**

plinth DATA ELEVATED

Makes sense for a network; expensive for one study

Choose it when

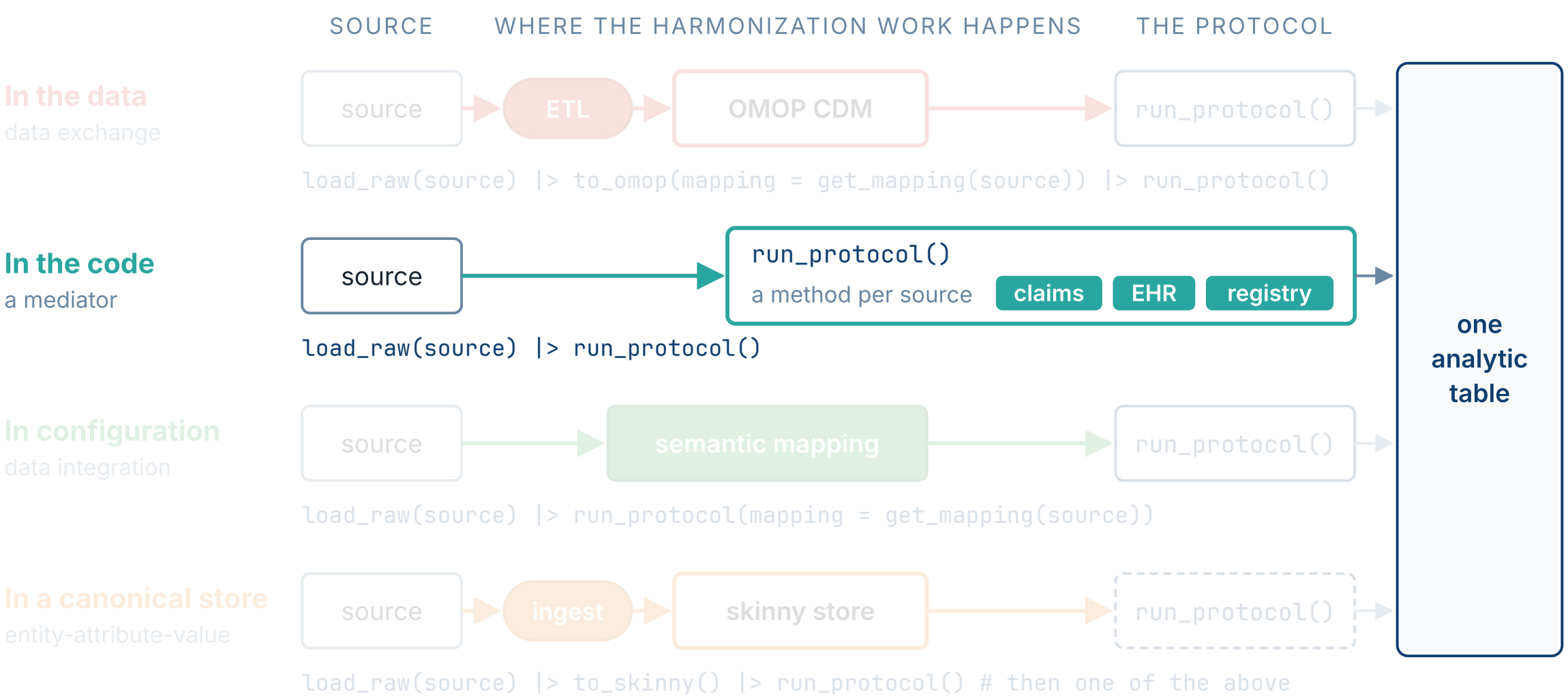
- a network needs **one query to travel** across many databases
- the definitions **do not change** from study to study

Pay for it in

- mapping the whole source for every future question: **97%** of the mapped rows are never read by this one
- an ETL validated as a whole; the study needs it validated for **its** codes

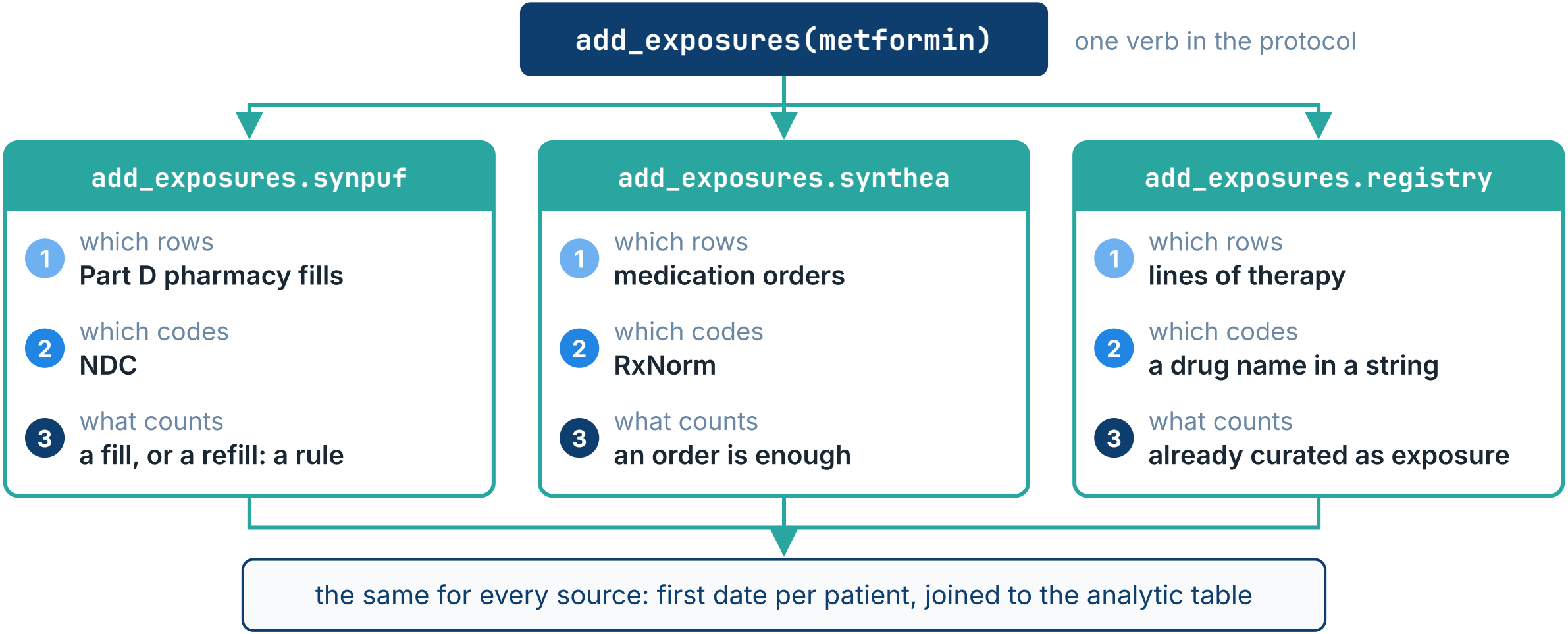
> Top-20 pharma datum: onboarding an already-standardized source ≈ **1 week**, a non-standardized one ≈ **5**.

In the code: a mediator. One interface, a wrapper per source, next to the question being asked



plinth DATA ELEVATED

One verb, a method per source: the science is a named function



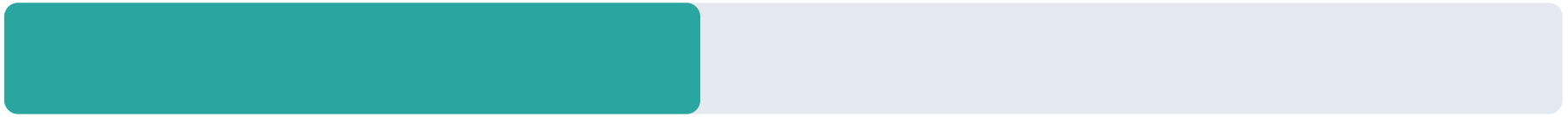
The exposure rule is an argument to the verb that owns it

`run_protocol()` any single fill counts as exposure



8,547

`run_protocol(min_fills = 2)` exposure needs a refill



3,815

patients in the claims cohort

The rule is a **parameter you can set, report and defend**, because it lives in a method with a name.

The most powerful and the most targeted, and the most code

Choose it when

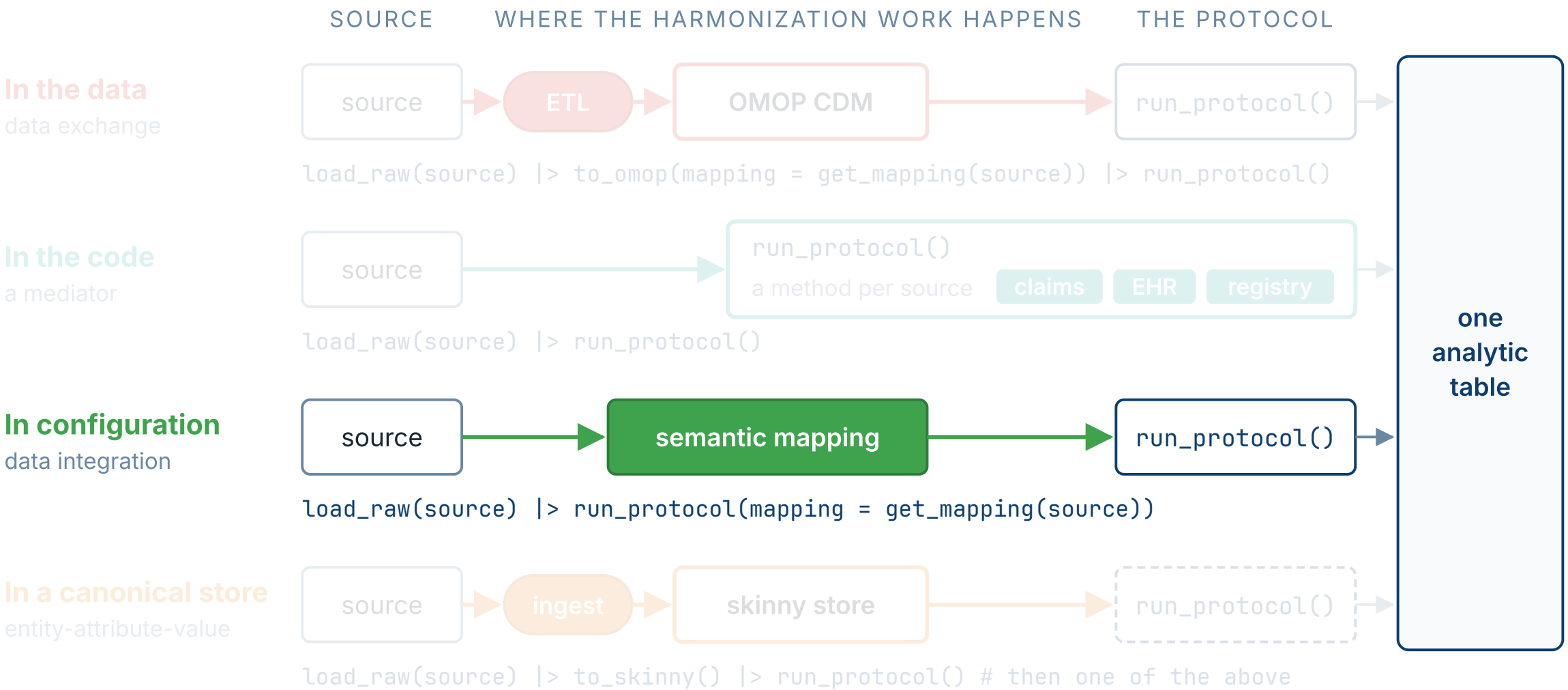
- the algorithm **differs by source**, or you expect several algorithms
- the science must be **reviewable** next to the verb that uses it

Pay for it in

- a method per verb per source
- no harmonized data product at the end

> Top-20 pharma datum: ~40 verbs × 11 datasets, ~200 methods, owned by the analysts who own the science.

In configuration: data integration. A declared mapping, and the query rewritten through it when it runs



plinth DATA ELEVATED

A mapping per source: where the evidence lives

THE MAPPING FILE, SYNPUF.YML

```
drug:
  table: prescription_drug_events
  patient_id: DESYNPUF_ID
  date: SRVC_DT
  code: PROD_SRVC_ID
  code_system: NDC   the vocabulary
```

THE CLAIMS SOURCE, AS IT ARRIVED

prescription_drug_events

DESYNPUF_ID	SRVC_DT	PROD_SRVC_ID
0A6F99CACE…	20091218	55289038494
10841025CC…	20080421	55289038494

Every line answers **where**. No line answers **what counts**.

plinth DATA ELEVATED

The verb reads the mapping: one function for every source

```
load_raw("synpuf", con) |> run_protocol(mapping = get_mapping("synpuf"))
```

```
add_diagnoses <- function(dataset, codes,
                          mapping = NULL, ...) {
  if (is.null(mapping)) UseMethod("add_diagnoses")
  dx <- mapped_events(dataset, mapping,
                      "condition") |>
    with_codes(codes)
  dataset |>
    join_analytic_table(first_date(dx, codes))
}
```

Given a mapping, the verb **reads the file** instead of calling a source-specific method.

Same file, same query, every time.

What the mapping holds, and what stays in the code

IN THE MAPPING FILE

Where the evidence lives

- which table, and the patient key
- which column holds the date, in what format
- which column holds the code
- which vocabulary, and which unit

A data engineer can write it from the codebook.

IN THE CODE, BESIDE THE VERB

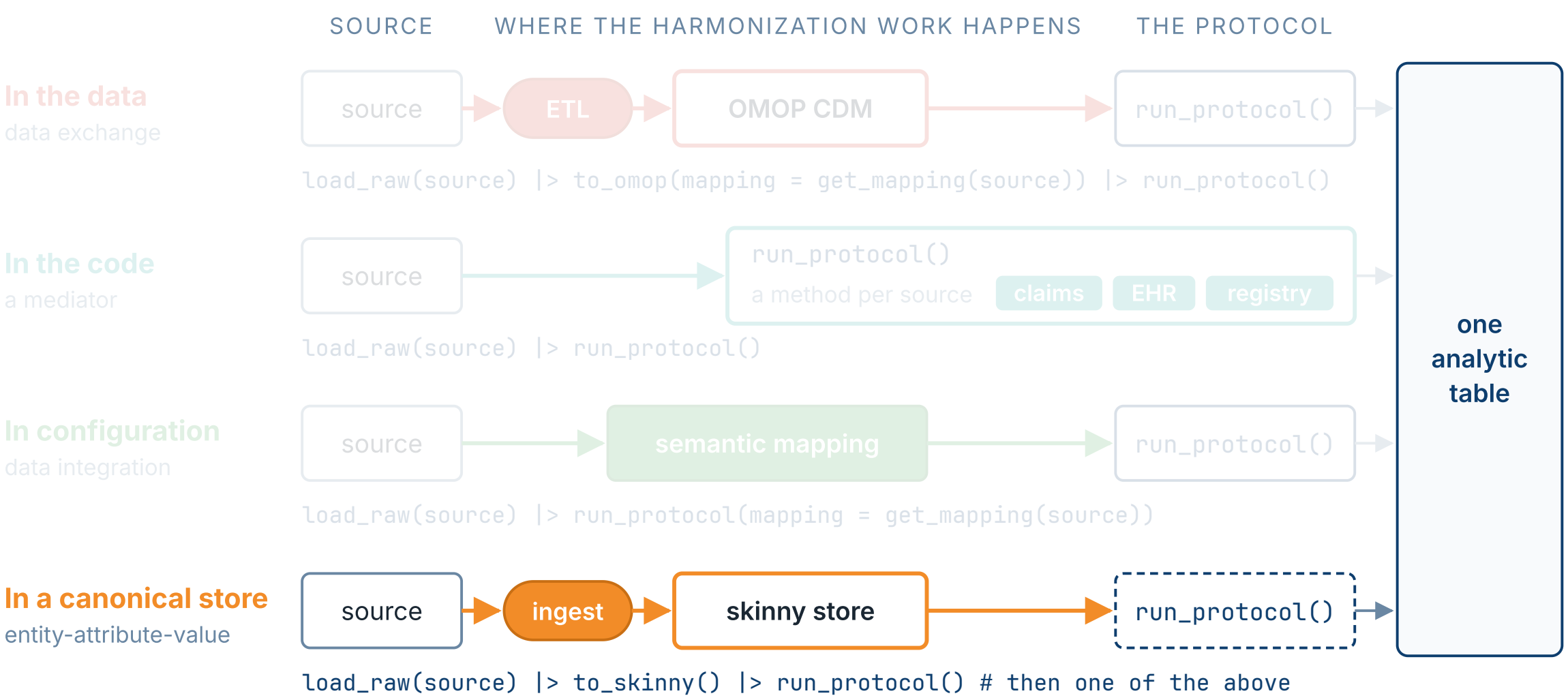
What counts as evidence

- which codes are metformin
- one fill, or a refill
- whether day 0 is inside the window
- who is in the cohort

A scientist decides it, and defends it.

The same function ran on every source. **Only the file changed.**

A canonical store: entity-attribute-value. Strip the schema, defer the meaning, then use one of the above



plinth DATA ELEVATED

A canonical store: every source into one long table

In: one claim, wide

DESYNPUF_ID	CLM_FROM_DT	ICD9_DGNS_CD_1	ICD9_DGNS_CD_2	ICD9_DGNS_CD_3
00256A69...	20080213	25001	2449	6256

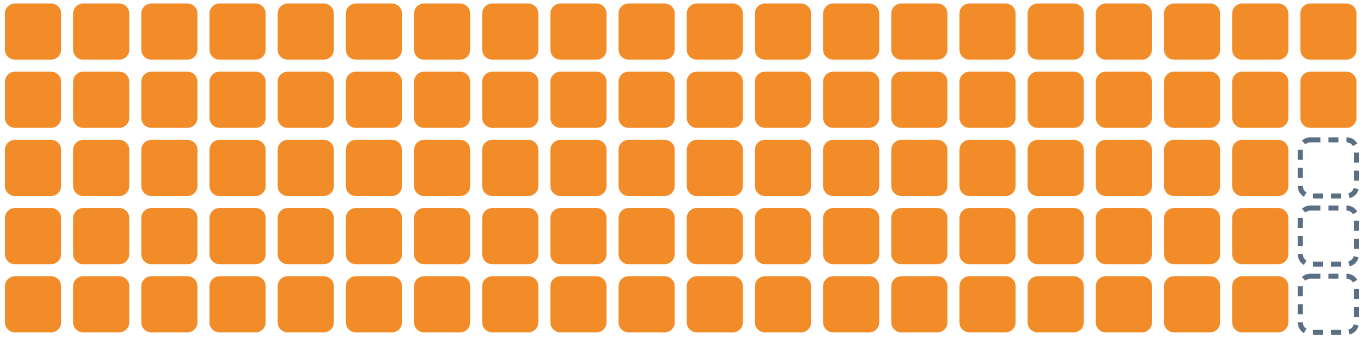
Out: one fact per diagnosis, long

entity_id	domain	code	code_system	event_date
00256A69...	condition	2449	ICD9CM	2008-02-13
00256A69...	condition	25001	ICD9CM	2008-02-13
00256A69...	condition	6256	ICD9CM	2008-02-13

The ingest reshapes. It maps **no vocabulary** and applies **no rule**.

What the store leaves to the query: the source’s vocabularies and units

```
results <- facts(dataset, "measurement") |>
  with_codes(lab) |>
  filter(unit == "%")        # the line that lost the mmol/mol rows
```



each square: 1% of the registry's HbA1c results

97% reported in %
kept by the filter

3% reported in mmol/mol
3,218 rows dropped, no error

Every decision the store deferred is now **every analyst’s decision**, made silently in a query.

They compose: a mapping over a store, the protocol over the CDM

```
raw <- load_raw("synpuf", con)
raw |> to_skinny() |> run_protocol()                          # store, then the code
raw |> to_skinny() |> run_protocol(mapping = skinny_mapping)  # store, then a mapping
raw |> to_omop(m)  |> run_protocol()                          # CDM, then the code
raw |> to_omop(m)  |> run_protocol(mapping = omop_mapping)    # CDM, then a mapping
```

Claims: patients

the protocol runs over	with the code	with a mapping
the raw source	8,547	8,547
a canonical store	8,547	8,547
the OMOP CDM	6,962	6,962

Registry: with a baseline HbA1c

the protocol runs over	with the code	with a mapping
the raw source	86.5%	86.5%
a canonical store	85.9%	85.9%
the OMOP CDM	86.5%	86.5%

Composition works, and it **carries the substrate's decisions with it**.

Where the six paths disagreed, and why

CLAIMS PATIENTS

8,547 → 6,962

2 interpret the values

The vocabulary mapping filed codes outside the concept the query asks for.

Report
the concept set

CLAIMS PATIENTS

8,547 → 3,815

3 what a record means

Whether one pharmacy fill is exposure, or exposure needs a refill.

Report
both rules, as a sensitivity analysis

REGISTRY, WITH A BASELINE

86.5% → 85.9%

2 interpret the values

The store kept the unit as a column, and the query ignored it.

Report
the units, and the rows excluded

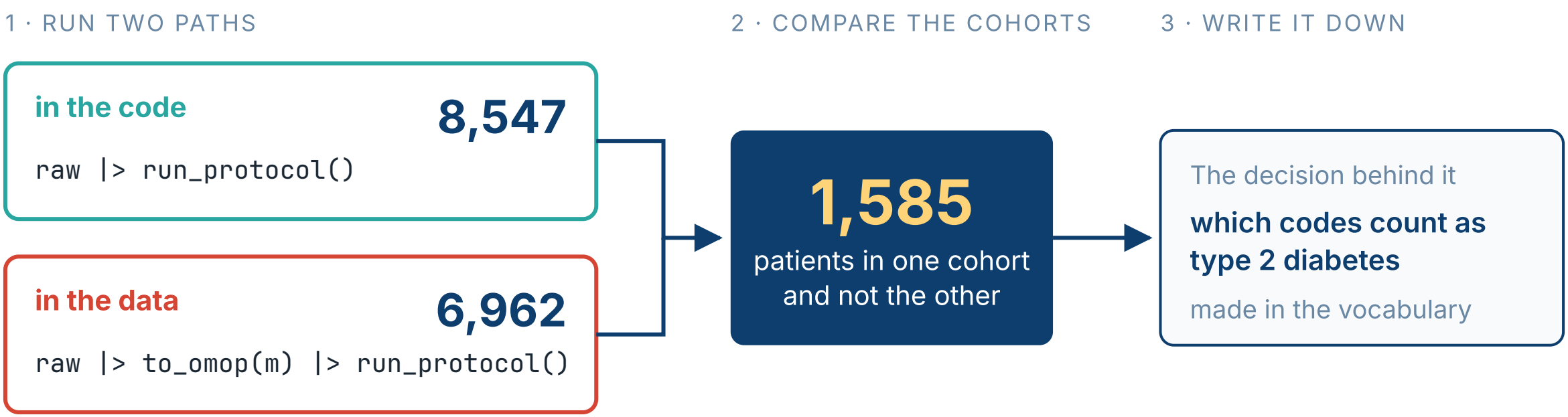
None of these is a property of a technique. **Each is a decision.**

One study, harmonized four ways: which fits the program you run

If you expect	harmonize	because
several algorithms per concept	in the code	the algorithm is an argument to a named method
one algorithm, many sources in their native formats	in configuration	the file says where; the verb says what counts
definitions that **do not change** across studies	in the data	one ETL, and one query that travels
questions not yet known	in a canonical store, then one of the above	the shape is solved now; the meaning is deferred

Choose by **how many algorithms you expect**, and **how early you know them**.

Harmonization does not remove decisions; it chooses where they are made



Keep the decisions visible. **Run two paths, compare the cohorts, and write down every difference.**

Kieran Mace · kieran@plinth.co · plinth.co

Backup

Backup: the ETL, in R

```r
etl_source <- function(con, mapping) {
  write_persons(con, mapping)                    # the roster the mapping names
  write_domain(con, mapping, "condition")        # every domain the mapping binds:
  write_domain(con, mapping, "drug")             #   the rows the protocol would read,
  write_domain(con, mapping, "measurement")      #   every code replaced by a standard concept
  write_observation_period(con, mapping)
}

write_domain <- function(con, mapping, domain) {
  if (!binds(mapping, domain)) return(log_unbound(con, mapping, domain))
  events <- mapping_rows(con, mapping, domain) |>  # the reader the protocol uses
    standardize(con, mapping, domain)              # the step only exchange needs
  events |>
    to_cdm_rows(con, mapping, domain) |>
    write_cdm(con, cdm_tables[[domain]]$table)
  log_step(con, mapping, domain, events)
}
```

`to_omop(mapping = )` runs this once per source. `mapping_rows()` is the reader the protocol uses too; `standardize()` is the step only exchange needs: source code to standard concept via *Maps to*, or a concept **asserted** from a label.

Backup: what the ETL wrote, a few rows of the CDM

condition_occurrence (SynPUF)

person_id	condition_start_date	condition_source_value	condition_concept_id	concept_name
3741847388539	2008-09-01	25000	201826	Type 2 diabetes mellitus
3741847388539	2008-05-23	25040	192279	Disorder of kidney due to diabetes mellitus
3741847388539	2008-12-25	4019	320128	Essential hypertension

drug_exposure (SynPUF)

person_id	drug_exposure_start_date	drug_source_value	drug_concept_id	concept_name
3741847388539	2008-05-19	55289038494	40163924	24 HR metformin hydrochloride 500 MG Extended Release Oral Tablet
3741847388539	2008-09-19	67046046130	40164946	metformin hydrochloride 850 MG Oral Tablet

Source value on the left, standard concept on the right. `250.40` (“diabetes with renal manifestations”) did not become a Type 2 diabetes concept. Remember that.

Backup: the ETL log

source	table	rows out	mapped %	decision
registry	condition_occurrence	10,000	100.0%	no source codes; concepts asserted from labels in the mapping
registry	drug_exposure	29,094	99.3%	local names -> RxNorm ingredients matched by name; unmatched -> concept 0
registry	measurement	226,520	100.0%	no source codes; concepts asserted from labels in the mapping
SynPUF	condition_occurrence	2,680,891	99.5%	ICD9CM -> standard concept via Maps to; unmapped -> concept 0
SynPUF	drug_exposure	5,552,641	99.6%	NDC -> standard concept via Maps to; unmapped -> concept 0
SynPUF	measurement	0	n/a	the mapping binds no such domain
Synthea	condition_occurrence	833,331	99.7%	SNOMED -> standard concept via Maps to; unmapped -> concept 0
Synthea	drug_exposure	1,164,220	100.0%	RxNorm -> standard concept via Maps to; unmapped -> concept 0
Synthea	measurement	6,131,943	100.0%	LOINC -> standard concept via Maps to; unmapped -> concept 0

Backup: a source can refuse

```
add_baseline.synpuf <- function(dataset, lab, ...) {
  not_available(dataset, "add_baseline",
                "claims carry no laboratory results")
}
```

The refusal is a **named failure**, not an NA column.

Backup: trade-offs, as a leave-behind

	in the data	in the code	in configuration	canonical store
Database name	data exchange, schema-on-write	mediator with wrappers	data integration, schema-on-read	entity-attribute-value
Where “what counts” lives	the ETL, once, upstream	a method per verb × source	the verb; the file holds only “where”	deferred to every query
Study-specific?	no (that is the point)	yes	can be	no
Onboarding a source	weeks to a year of FTE for a full CDM; paid once	one method per verb the study uses	one file	minutes
Work this question never uses	most of it (97% of mapped rows here)	none	little	none now, all of it later
Audit artifact	the ETL log, if kept	method source + attrition trail	file version + the query it drives	the fact row
Schema drift	re-run ETL	edit one method	edit one file	re-ingest
Data drift (new unit, new code system)	re-map, re-validate	one method, tested	one line	silent until a query notices
Who maintains it	data engineering	analysts who own the science	whoever owns the map, signed against the codebook	data engineering

Backup: the attrition trail is written by the verbs

```
<cohort> source: synthea  path: raw
 step                    n
 load                    23046
 add_diagnoses.synthea 23046
 add_exposures.synthea 23046
 add_index_date          935
 add_age                 935
 restrict(age >= 18)     935
 add_baseline.synthea    935
```

Evidence verbs annotate (N holds); `add_index_date()` and `restrict()` filter. The trail travels with the cohort on every path.

Backup: join on the code, not the label

DESCRIPTION	chars	rows	patients
Hemoglobin [Mass/volume] in Blood	33	75271	23046
Hemoglobin	10	51	51

Two descriptions for the same LOINC code, one of them a bare `Hemoglobin`. A join on the label silently drops the second group; a join on the code does not. In the previous export of this dataset the two labels differed only by a zero-width space.

Backup: codelists

```
List of 3
 $ t2d      :List of 6
  ..$ name           : chr "t2d"
  ..$ label          : chr "Type 2 diabetes mellitus"
  ..$ icd9           : chr "250"
  ..$ snomed         : chr "44054006"
  ..$ omop_concept_id: int 201826
  ..$ local          : chr "type 2"
 $ metformin:List of 6
  ..$ name           : chr "metformin"
  ..$ label          : chr "Metformin"
  ..$ omop_concept_id: int 1503297
  ..$ rxnorm_code    : chr "860975"
  ..$ ndc_table      : chr "vocab.metformin_ndc"
  ..$ local          : chr "metformin"
 $ hba1c    :List of 5
  ..$ name           : chr "hba1c"
  ..$ label          : chr "Hemoglobin A1c"
  ..$ loinc          : chr "4548-4"
  ..$ omop_concept_id: int 3004410
  ..$ local          : chr "hba1c"
```